



下载APP



01 | 爬虫需求的诞生：我们是正经的软件工程师

2022-01-19 DS Hunter

《反爬虫兵法演绎20讲》

课程介绍 >

**讲述：DS Hunter**

时长 16:15 大小 14.89M



你好，我是 DS Hunter，反爬虫专家。

也许你是一个爬虫工程师，也许是反爬虫工程师，甚至，也许你只是一个业务方的普通研发，被授予了反爬的重任。但是，不论你的身份是什么，“什么是爬虫”这个问题都是你必须了解的。

为什么这么说呢？

可能你常把爬虫挂在嘴边，觉得自己已经很熟悉爬虫了，但当你尝试自己做一个爬虫，完成一个反爬虫动作时，却发现无从下手。其实，很大的一个原因就是对于“什么是爬虫”这个问题了解得并不透彻。



从历史的视角来了解爬虫从哪里来、能做什么，以及从诞生到现在的这段时间里都发生了什么样的变化，可以让你对“什么是爬虫”这个问题产生更深度的思考，这也是我要在课程里特地为你设置一个“历史背景篇”的主要原因。咱们接下来正式开始吧。

什么是爬虫？

爬虫是一个历史悠久的需求，严格来说，它甚至比网络出现得还早。或者你也可以理解为，网络出现之后，网络和爬虫才结合成了我们所熟知的网络爬虫。因为互联网大部分的功能其实并没有什么新意，只不过是把线下的场景搬到线上来了。

而爬虫，其实就起源于线下。再聚焦一些，爬虫，起源于再平常不过的——菜市场。

买菜和爬虫？买菜也爬虫？

前几天我听了一首很喜欢的歌，叫《说走就走》。里面有一段话，说的是：“走世界，看精彩，从 18 岁讲到现在，最后到巷口去买菜。”

本意是说，年少充满理想，长大后被生活压垮，每天沉迷于菜场买菜这种小事，再也没有了理想。但是实际上，**买菜并不是小事，它充满了技术含量。**

菜场买菜的大妈们砍价非常厉害——我感觉我这辈子都学不来。对她们来说，即使只差一毛钱，也可能会让她们选择去别的摊位买菜。也许你不太理解这种行为，觉得，我都在这问了价了，就因为隔壁少一毛钱，就去隔壁，这多不好意思啊。但是对于菜场的顾客来说，这很正常啊，“同样的东西，哪里便宜我就去哪呗”，这是再正常不过的“**博弈**”了。

说回喜欢在网上买东西的你。一个小东西，A 网站比 B 网站便宜一毛钱，你是不是就果断抛弃了 B 网站？或者 A 说，价格一样，我包邮。你是不是就直接去 A 网站买了？这个时候，因为没有面对面的交流，你就没什么不好意思的了。所以说，趋利避害这是人性使然，自古如此，与年代性格都无关。**只要有人的地方，就会有“博弈”。**

最后，在买菜或者买东西的故事里，除了“你”这个主角，商贩们其实也没闲着。他们也会做一些操作，比如让自己的家人去四处转转，看看别的摊位卖多少钱、有没有偷偷降价。如果有顾客来自己的摊位买菜，也会顺便问问，“哎你这个菜多少钱买的啊”，只要不低于成本，他就敢降到一样的价格，甚至更低。

实际上，大妈们获取价格的方法、你对比价格的过程以及商贩们相互获取价格的方法，都可以理解为“爬虫”行为，和网络爬虫其实也没有什么区别。只是人工问价效率低，爬虫效率高。价格，就是在这样不断博弈的过程中，慢慢均衡下来的。而博弈的第一步，就是“问价”——**获取数据**。

不止买菜：我只是要数据

为什么获取数据这么重要？其实《孙子兵法》就提到过：知彼知己者，百战不殆。商场如战场，获取数据自古就是胜负的关键。

至于获取数据的需求是怎么产生的，我给你举个例子吧。获取数据的手段——爬虫，很难学的一个主要原因就是描述它的词汇太多了。你可以看下网络上对爬虫的定义，有几个常用的词：爬虫、抓取、spider、crawler.....可能我也说不全。日常使用的话，这些词你随便挑哪一个都行。但是搜索对应知识的时候，为了更全面一些，就要全搜一遍。那如果你不想连搜四次，怎样操作才能**一次性拿到所有的结果**呢？

这个时候，聪明的你可能已经想到了，我要不要写个爬虫全拿下来呢？恭喜你，你为了获取爬虫的知识，已经自发地产生了一个爬虫的需求。

因此你会发现，**爬虫需求的产生是自然而然的**，而你的需求不一定是一个坏的需求——你看，我们多正经。是的，技术只是工具，与善恶无关。

最早的 Google 工程师就是这么想的，或者再早的 Yahoo。我们无从得知当时具体发生了什么、第一个爬虫是如何被写出来的。但可以想象的是，**因为当时的互联网数据越来越多，获取信息变得越来越难，于是他们就写了一个网络爬虫来获取数据**。从此，潘多拉的魔盒就被打开，后面的事情就不再可控了。

搜索引擎和爬虫

既然提到了 Google 和 Yahoo，我们就紧接着从历史视角，审视一下搜索引擎和爬虫的关系，看看爬虫的早期历史和相关的技术变迁。

美好的上古田园时代

搜索引擎可以认为是爬虫的集大成者了。它本质上就是用爬虫**爬取天下数据**，汇聚在自己的网站上，让大家可以在自己的网站上直接**搜索到全天下的知识**。这个操作对于搜索引擎

和站点两方都是有利的，搜索引擎自己能够获得稳定客户，甚至商业利益；站点本身可以获得流量，未来也可以获得商业价值。

为了这个双赢的局面，还引发了一个叫 robots.txt（也叫 robots 协议）的君子协定，里面会约定，哪些可以爬，哪些不可以爬。但是因为大家很依赖搜索引擎，所以除非是用户信息，否则一般都是开放的。甚至有些站点不写 robots.txt，默认对搜索引擎全部开放。除此之外，还可以写 sitemap 来指定自己网站都有什么，欢迎搜索引擎来赶紧爬走，给自己打个好评，引导用户过来。大家主动交流，相互沟通，真的是一个美好的上古田园时代。

春秋时期

但是美好的上古田园时代由于网络的不断发展，注定不会持久，很快，爬虫就发展到了春秋时期。

春秋前期

在春秋前期，爬虫已经不是新技术了。你的那个不懂技术的老板，可能都已经知道了：想要寻找大量数据，可以用爬虫去爬别人。所以，你就接受了这样的需求，开始学习相关的知识。

但是你也知道，网络历史从 TCP 到 HTTP，现在又回到了 TCP。随着历史的变化，爬虫本身的技术也会随之变化。毕竟，**服务器使用了什么类型的网络，爬虫就要被迫使用什么类型的网络。**

春秋中期

当爬虫开始受网络发展的影响，就注定也会被其它因素影响。在春秋前期，大家逐渐产生“爬取数据”的需求，到了春秋中期，这些需求逐渐演变成了一个业务的雏形。也就是说，在这个时候，**不同的业务开始对爬虫技术产生影响了。**

咱们先说**扒站**这件事。最初爬虫都是一些大公司的特权，小公司自己的业务都做不完，哪有时间去搞爬虫。但是出于业务的需要，一个新的行为就诞生了——站点复制，俗称扒站。

也就是说，别人网站做得好，我刚创业，追赶别人根本来不及。怎么办呢？答案很简单，我写个爬虫把整个网站抓下来不就好了。然后把数据放在我的服务器上，就可以瞬间追上别人的进度，站在同一起跑线。

还有就是**浏览量的需求**了。站长很想让搜索引擎来爬自己的数据，因为有了搜索引擎的爬取，就会有排名，就会有业绩。而为了提高搜索引擎的分数，他们十分渴望搜索引擎来抓取自己所有的数据。

关于搜索引擎爬取自己数据的这件事，我们可以回顾一下历史。

早年的网站结构其实很简单，就是一个服务器，上面挂了一些静态的文件。有的站点甚至会打开目录遍历权限，也就是说你去找一个目录看一下，能直接看到目录下面有什么文件，这种站点扒起来简直太舒服了。但是如果没有开目录遍历权限，就要麻烦很多了，似乎抓起来就没有前面说得那么容易。

好了，从历史回顾里跳出来，春秋中期的站长，为了给搜索引擎抓取数据提供便利，通常会有两种操作：一种是 **sitemap**，也就是给整个站点建一个地图，给对方使用；还有一种就是**内链**，通过自己不断引用自己，来实现引导爬虫爬取完整站。

这样一来，爬虫工程师的爬虫思路就显而易见了：既然站长提供了 sitemap 和内链两种便利，我只要声称自己是搜索引擎就好了。这样对方不但不会拒绝，还会引导我去爬取整个站点。这样，想扒整个站就变得非常容易。

你可能会觉得，这里有点引狼入室的感觉了——难道站长已经有准备了？没有。等狼真的来了也没有什么办法——撑死就是封一封明显太过分的 IP。

最后我来给你总结一下这个时期爬虫的爬取思路吧。春秋中期，爬虫基本上就两个爬取思路，一个是**深度优先遍历**，一个是**广度优先遍历**，它们的最终目的都是抓完整站。所以爬虫会有一个“spider”的名字，蜘蛛能结网，指的就是这个遍历方式。如果你以后面试的时候，面试官着重考查这两个知识点，你要理解，他已经很久不做爬虫了。这是远古时代的基本技能，现在大家已经不太关注这个了。

春秋后期

后来两件事情改变了这个行业的格局。一个是**电子商务的出现和普及**，一个是**站点的动态化和伪静态化**，我们也就随着这两件事从春秋中期进入到了春秋后期。

先看第一件事：电子商务的出现和普及。你还记不记得我在开头提到的一句话：互联网大部分功能其实并没有什么新意，只不过是把线下场景搬到线上来了。

电子商务其实就是把菜场搬到了线上。 菜场有爬虫，电子商务就会有爬虫。与搜索引擎不同的是，商场如战场，战场上的较量，有时候就没有那么强调道德了。爬虫也就渐渐**为所欲为**了起来。

另一件事就是站点的动态化和伪静态化，动态化页面导致站点内容变得十分丰富，甚至可以认为近乎无穷无尽，通过改变参数就可以不断改变站点内容，这一点对搜索引擎非常不友好。但是好消息是，这一点对其它爬虫，一样很不友好。

因此，站点开始进行伪静态化。假装自己依然是个静态站点后，爬虫的需求开始变得复杂，单纯的深度优先和广度优先满足不了大家的要求了。尤其是随着 Web 2.0 的诞生，站点更加复杂，本身也开始了分层，界面是界面，数据是数据。那么我们的爬虫最关注的是什么？是数据。那就是说，爬虫甚至可以只抓数据，跳过界面。这个时候，爬虫就变得**十分高效**。

不过，当爬虫变高效之后，服务器的噩梦就诞生了。由于前后端分离，静态资源放置于 CDN，通常并不是很怕爬虫。但是数据位于服务层，服务层比 CDN 脆弱得多，爬虫一来，可能**打爆链接数**，甚至可能**击穿数据库、拖慢站点性能**，各种诡异的情况都可能发生。

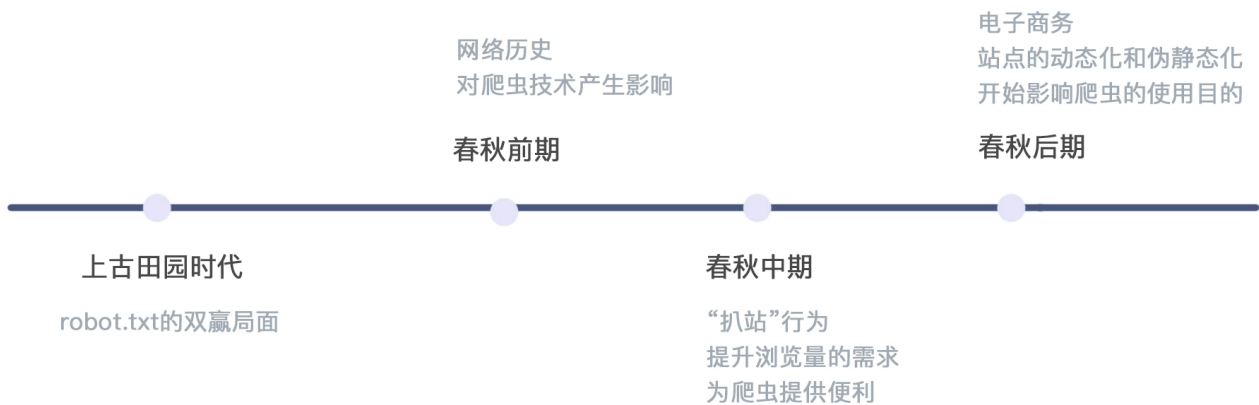
总的来说，**电子商务的出现和普及**以及**站点的动态化和伪静态化**这两件事出现之后，爬虫就很容易不小心惹事。很多文章都会教你，尽量控制频率，不要把站点爬挂。而我们被爬的站长那一方，他欢迎搜索引擎，但是不喜欢爬虫，不过依旧没有太好的办法区分彼此。

如果说爬虫是一场战争，那么现在的春秋时代，大家打仗还是讲究一个“礼”字的。截止到目前，还没有彻底崩坏。但是，根据历史我们可以知道，这样的事情不会持续多久，战国，很快就要来了。

小结

好了，最后我来给你小结一下。今天和你聊了爬虫的产生、早期历史以及一些相关的技术变迁。

可以说，“爬虫”是竞争的必然产物，而网络的出现，赋予了爬虫在互联网领域的生命。再次强调，这项技术只是工具，与善恶无关。



在美好的上古田园时代，爬虫彬彬有礼，用技术不断辅助大家，让整个互联网变得更好。搜索引擎自己能够获得稳定客户，甚至商业利益。而站点本身可以获得流量，未来也可以获得商业价值。你看，**需求正经，做的也是正经事。**

截止到这个时候，爬虫还没有任何过错，还没有到现在这种人人喊打的地步。如果人类的爬虫技术止步于此，那么这个世界将非常美好。遗憾的是，人性自古不变，行业建设到一定程度就会产生内卷。

在春秋时期，随着不同业务需求的诞生，爬虫技术也随之发展。春秋前期，我们发现爬虫开始不停地迭代，适应网络环境。后来，春秋中期的扒站行为，本身已经是一种轻度内卷了。同时，站长出于对浏览量的需求，提供了 sitemap 这张地图以及内链的方式，助长了爬虫疯狂爬取的气焰。而到了春秋后期，电子商务更为爬虫增添了不少的商业气息，商人逐利，爬虫也逐渐为所欲为。站点的动态化和伪静态化，让爬虫直接爬取数据层，站点无法承受攻击……

下一讲，我们会进入到战国时期，看下礼乐制度彻底崩溃、内卷到血流成河的时候，整个行业会成为什么样子。而这，也是你我共同面临的环境。

思考题

好了，这次是我第一次给你留思考题。下面有三个方向，你可以选择一个来和我分享：

1. 站长在喜欢搜索引擎的同时痛恨爬虫，不过搜索引擎本身也是一种爬虫。那么，假如一个爬虫冒充搜索引擎，怎么办？
2. 爬虫为这个世界做出了什么贡献？
3. 你的爬虫或者反爬虫经历是什么？有什么奇葩的经历吗？

期待你在评论区的分享，我会及时回复，不过要记得注意保密脱敏。

“

混沌未分网路乱，圈地既成爬虫生。



反爬虫兵法演绎 20 讲

DS Hunter / 反爬虫专家，前某公司研发总监

”



识别二维码
免费试读



分享给需要的人，Ta订阅超级会员，你将得 50 元

Ta单独订阅本课程，你将得 20 元

 生成海报并分享

 赞 3  提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 [开篇词 | 如何突破“爬虫反爬虫”内卷之怪现状？](#)

下一篇 [02 | 爬虫的内卷和黑化：我们变得不正经啦](#)

精选留言 (4)

写留言



程序员二师兄

2022-01-20

本人刚接触爬虫这一块不久，看到思考题留的作业，按照自己对爬虫浅薄的理解回答一下：

1. 假如爬虫冒充搜索引擎，怎么办？

...

展开

作者回复: 哈哈哈，还抓过搜索引擎，可以。其实搜索引擎并不好抓，只是抓的少的时候不触发任务反爬处理，它们不当回事。



1



fsc2016

2022-01-20

感觉现在爬虫对抗，慢慢从web转战到移动端了

作者回复: 是的，移动端如果感兴趣，后续可以再出课程



peter

2022-01-20

写得太好啦，牛啊，文采真好，通顺，有趣，也没有错别字！！！！我还有很多课程没有看，这个课程和自己目前的关系并不是很大，犹豫再三才买的。真庆幸自己没有错过这么好的文章。理工男能写这么好，不容易，不多见啊。

作者回复: 太感动了，泣不成声。





GAC·DU

2022-01-19

一次经历见证了一家公司技术的成长，甚至把后端由15人加到了30。从携cookie能登录到手机验证码再到扫脸登录，api加了token，后来把限流和熔断也加上了。

作者回复: 然而还是有可能无效对吧，最关键，有效无效还很难验证。。。除非把对方收购了直接问他🤖🤖🤖

